**Unsupervised learning**

**Introduction of the unsupervised learning**:

Unsupervised learning is a method of machine learning, which can automatically classify or group the input data without giving the training examples. The main applications of unsupervised learning include **cluster analysis**, **association rule** and **dimension reduction**.

**Recommendations for a person who want to develop or use such systems:**

State of the art un-supervised learning method:

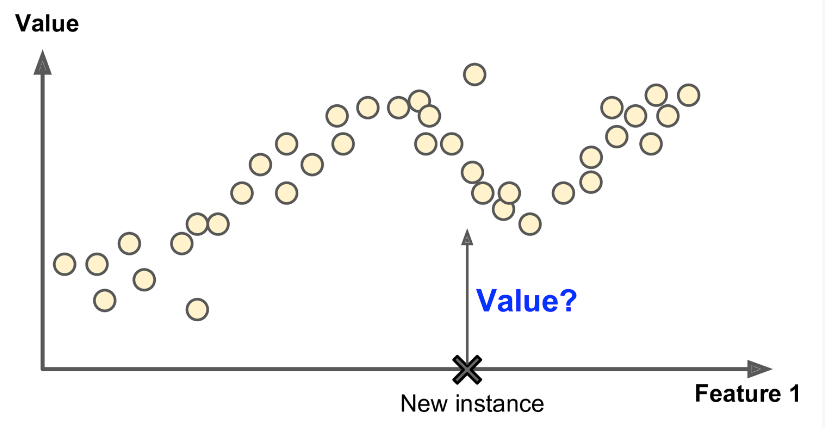
**K-means clustering**: A method originated from signal processing is more popular in the field of data mining as a clustering analysis method. The purpose of K-means clustering is to divide n points into K clusters, so that each point belongs to the cluster corresponding to its nearest mean, and take it as the standard of clustering. This problem is very difficult in computation, but there are efficient heuristic algorithms. Generally, efficient heuristic algorithms are used, which can converge to a local optimal solution quickly. These algorithms are usually similar to the maximum expectation algorithm which deals with Gaussian mixture distribution by iterative optimization. Moreover, they all use clustering centers to model data; however, K-means clustering tends to find clustering in a comparable spatial range, while expectation maximization technology allows clustering to have different shapes.

**Mixture model**: It is a probability model used to represent the existence of the population.

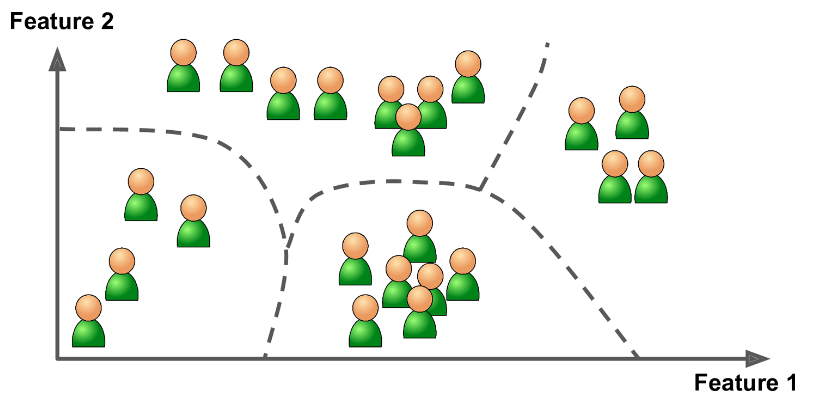
**Anomaly detection**: Identification of items, events, or observations that do not conform to the expected pattern or other items in the data set. Under the assumption that most of the instances in the data set are normal, the unsupervised anomaly detection method can detect the exceptions of unmarked test data by finding the most mismatched instance with other data.

**Pros and Cons: Compared with supervised learning:**

In supervised learning, there are labels. They are the expected results in the algorithm. The common tasks are classification. Like the spam filter, the filter gets a e-mail, and it will know whether it belongs to normal mail or spam.



Training data for unsupervised machine learning is not labeled. The system will try to learn by itself. For example, if you have a website, you have a lot of data about your website visitors. You can use clustering algorithm to try to group similar visitors, and the algorithm will find the association between visitors. For example, it may find that 40% of the visitors are male. They like comics and usually read your articles at night. 20% of them are young science fiction fans and often visit your website on weekends.



**Conclusions:**

Unsupervised learning is very important to understand the change and grouping structure of a set of unlabeled data. It can be used as a useful processor of supervised learning. It is essentially more difficult than supervised learning because there is no gold standard and a single goal. It is an active research field with many recently developed tools, such as self-organizing mapping, independent component analysis and spectral clustering.

**Summary of teammates' report:**

Dingjuan Dian:

This report is about introduction the three major directions in machine learning, large scale machine learning ,deep learning and reinforcement learning. In deep learning, he introduced Siri's voice.

Suli Hu:

He introduced that Unsupervised learning is a type of self-organized Hebbian learning that helps find previously unknown patterns in data set without pre-existing labels. It is also known as self-organization and allows modeling probability densities of given inputs.auto August from Google, which is a deep reinforcement learning method to automatically enhance training data.

**Reference:**

[1] Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron

[2] <https://ninghao.net/blog/5607>

[3] <https://en.wikipedia.org/wiki/K-means_clustering>